知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【工具方法】

词共现频次变化视角下的动态主题识别研究

席崇俊 刘文斌 丁楷
中国科学技术信息研究所 北京 100038

摘要：[目的/意义]主题识别研究对于理清领域内的知识结构与研究热点非常重要，对领域主题进行动态识别，可以很好地帮助研究人员了解和掌握领域的发展态势及未来走向。[方法/过程]利用张量的数据结构形式，在词共现矩阵中融入时间维度，只需一次聚类便可进行动态主题的识别。[结果/结论]张量结构及非负张量分解算法为词共现频次变化视角下的动态主题识别提供一种新的方法，该方法相较于传统方法更为简单快捷，有效避免了信息的损失。

关键词：关键词共现　非负矩阵分解　非负张量分解　动态主题识别　知识管理
分类号：G254.2

引用格式：席崇俊，刘文斌，丁楷．词共现频次变化视角下的动态主题识别研究[J/OL]. 知识管理论坛，2022, 7(2): 197-208[引用日期]. http://www.kmf.ac.cn/p/281/.

1 引言

在信息时代背景下，随着科技文献数量的迅猛增长，研究人员无法在短时间内吸收和掌握数以万计的研究成果，即便是针对范围狭窄的领域进行密切关注、持续阅读，仍难理清该领域的研究热点和研究方向[1]。因此，对领域主题的挖掘与演化研究则显得尤为重要，它可以很好地帮助研究人员了解和掌握领域的发展态势及未来走向，也是解决信息大爆炸时代情报危机的有效方法[2-3]。本文基于词共现频次变化视角对动态主题识别方法进行探讨，旨在为科技决策提供更好的支持。

2 研究现状

主题识别与演化研究是利用文献特征项之间的关联关系对文献集合进行分析从而发现主题，并通过主题揭示文献集合中蕴涵的内容，以了解当前领域的研究热点并预测未来的发展趋势[4]。在主题识别与演化分析研究中，相关学者已经开展了大量研究，根据研究对象由浅及深可分为基于文献外部引用关系的方法、基于文献内部词分析的方法、基于全文内容文本挖掘的方法等。

基于文献引用关系的分析方法可分为文献共被引法、文献耦合法以及文献间的直接引用法等，主要是利用文献之间的引用关系来判断文献之间的关联程度，从而对文献进行划分，达到主题聚类的目的[5-6]。例如祝清松等提出基于引文主路径文献共被引的主题演化分析方法，通过对引文主路径上关键文献的共被引分析来

作者简介：席崇俊，硕士研究生，E-mail：xicj7465@163.com；刘文斌，硕士研究生；丁楷，硕士研究生。
收稿日期：2021-10-22　　发表日期：2022-03-24　　本文责任编辑：刘远颖

揭示学科领域的主题演化情况 [7]；黄福等通过核心文献与其被引文献进行耦合分析，再通过核心文献及其施引文献进行共被引分析，进而分别构建研究前沿领域 [8]；宋艳辉等以 SCI 和 SSCI 收录的 7 种情报学期刊在 2000-2010 年间的数据为样本，以作者文献耦合分析方法为研究视角，探寻新世纪以来情报学的知识结构 [9]。

基于词分析的方法主要分为词频分析法和词共现分析法，词频分析法是通过统计文献中关键词出现频次的高低变化来确定领域的研究重点及热点 [10]，词共现分析法则是通过统计一组词共同出现的次数来分析词之间的关联关系，从而对词进行聚类得到主题 [11]。例如奉国和等基于生命周期理论和词频分析方法，对学科领域发展过程进行客观合理的动态跟踪与分析 [12]；储节旺等运用词频分析法，通过对文献关键词的词频统计，进而对近 10 年来知识管理领域的研究热点、应用领域和研究方法进行分析 [13]；姜鑫等利用 CNKI 数据库通过词频分析法结合共词分析法对 2005-2016 年我国科学数据领域的研究主题进行演化分析 [14]；赵丽梅等以共词分析为基本研究框架，揭示大数据背景下数字图书馆研究领域的主流研究范式，为后续研究提供内容基础和理论依据 [15]；唐果媛等采用人工判读法提炼出基于共词分析法的学科主题演化研究分析流程的 5 个步骤，并对每个步骤中研究人员使用的策略、分析手段和工具进行归纳总结 [16]。

基于文本挖掘的方法则是通过文本挖掘技术对主题进行抽取，并用相关评价标准对主题进行分类。例如胡吉明等构建了适用于动态文本内容主题挖掘的 LDA 模型 [17]；杨超等构建了基于“主语—行为—宾语”（subject-action-object, SAO）结构的 LDA 主题模型，实现对专利文献主题结构的识别和分析 [18]；J. Kim 等通过文本挖掘和决策树的方法进行技术预测，从论文作者、期刊、所属领域及专利的专利权人、所属领域等字段中抽取能代表技术主题领域的特征 [19]。

其中，基于词共现分析的方法可以深入到文献内部，既关注词出现的频次大小，也考虑了词间的语义关系，是当前较为广泛使用的一种方法。因此，本文考虑基于词共现的分析方法对领域主题进行挖掘。传统基于词共现分析对多个周期的主题进行动态识别时，通常是基于二维数据——要么是根据各年份的词频变化矩阵进行聚类；要么是先按年份对词进行时间切片，然后分别构造词共现矩阵进行单独多次聚类，从而实现动态主题识别。前一种方法未考虑词间的语义关系，后一种方法则需要进行多次聚类，损失了大量信息。本文考虑借助张量的数据结构形式，在词共现矩阵上融入时间维度，构造三维数据，并基于非负张量分解算法只需一次聚类便可得到各年份的主题情况，有效减少了数据的损失。

3 研究思路

本文的具体研究思路如图 1 所示：

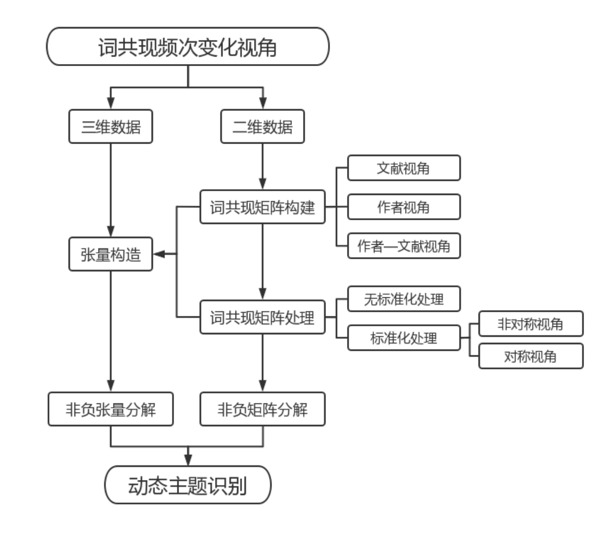


图 1 研究框架

为了在词共现矩阵中融入时间维度，从词共现频次变化视角下进行动态主题识别，本文首先对词共现矩阵的构造方式、数据处理方式以及聚类方法进行探讨。①词共现矩阵的构建。文献是关键词的载体，而作者是科学研究

知识管理论坛，2022（2）:197-208
DOI: 10.13266/j.issn.2095-5472.2022.017

的主体，二者所使用的关键词集合对领域的知识结构有着不同的反映，因此，本文考虑分别从文献视角和作者视角构建关键词共现矩阵，并将两种视角下的矩阵进行融合，比较基于三种关键词共现矩阵得到的主题识别结果的差异。②词共现矩阵的处理。在基于共现数据进行研究时，有学者指出直接在原始数据上进行分析即可 [20]，有学者则认为需要对原始数据进行标准化处理后再进行分析 [21]，在以往基于关键词共现的主题识别研究中，关于共现矩阵是否需要以及如何进行标准化处理尚无统一定论，因此，本文分别从对称视角和非对称视角对关键词共现矩阵进行标准化处理，对比是否进行标准化操作以及不同的标准化处理操作方式对主题识别结果的影响。③词共现矩阵的聚类方法。非负矩阵分解算法相较于传统聚类算法（系统聚类法、主成分分析、奇异值分解等）可以有效避免关键词与类团的单属性以及权重值为负等不足，而非负张量分解是非负矩阵分解在高维空间的拓展，因此，本文首先明确非负矩阵分解算法相对传统聚类算法的有效性，然后比较非负分解算法与非负张量分解算法在动态主题识别中的优劣性。

4 数据集及研究方法

4.1 数据集

4.1.1 数据集的构建

本文在 Web of Science 数据库中以“knowledge management”为主题词检索了国外知识管理领域相关文献，文献类型限定为“article”，文献时间为“2017-2021 年”，共检索到 4 898 篇文献，包含 11 343 个关键词字段和 12 178 个作者字段，通过对数据字段进行清理，去除本位词“knowledge management”的影响，选择频次大于 1 的关键词进行研究，并按如下三种方式构建本文所需的关键词共现矩阵：

（1）文献视角下的关键词共现矩阵构建。假设 $KT^{m\times p}$ 为关键词—文献共现矩阵，其中 m 为关键词数，p 为文献数，矩阵元素为关键词在文献中出现的次数，显然 $KT^{m\times p}$ 为 0-1 值矩阵，则基于文献的关键词共现矩阵 $AT^{m\times m}$ 可定义为：

$$AT^{m\times m}=KT^{m\times p}*(KT^{m\times p})^T \qquad 公式（1）$$

（2）作者视角下的关键词共现矩阵构建。同样地，假设 $KR^{m\times q}$ 为关键词—作者共现矩阵，其中 m 为关键词数，q 为作者数，矩阵元素为作者使用关键词的次数，则基于作者的关键词共现矩阵 $AR^{m\times m}$ 可定义为：

$$AR^{m\times m}=KR^{m\times q}*(KR^{m\times q})^T \qquad 公式（2）$$

（3）融合文献和作者双视角下的关键词共现矩阵构建。考虑到无论是基于文献还是基于作者的关键词共现本质上都是计算关键词共同出现的次数，区别在于一个从文献视角考虑，一个从作者视角考虑。对同一个领域来说，某一时间段内其所包含的研究成果是一定的，由于科技文献是研究成果的载体，而作者是科学研究的主体，二者互为补充，从不同视角对领域内的研究情况进行了划分，因此本文考虑同时结合这两个视角，融合文献和作者的关键词共现矩阵 $ATR^{m\times m}$ 可定义为：

$$ATR^{m\times m}=AT^{m\times m}+AR^{m\times m} \qquad 公式（3）$$

4.1.2 数据处理

（1）对称视角下的标准化处理。2009 年，N. J. van ECK 等指出在对共现数据进行分析时需要利用相似性度量来标准化数据，并对比了几种常用的相似性度量方法（关联强度、余弦相似度、包含指数、Jaccard 指数），发现基于概率的相似性度量方法（关联强度）效果要好于基于集合论的度量方法（余弦相似度、包含指数、Jaccard 指数）[22]。因此，本文将利用关联强度计算公式对关键词共现矩阵进行标准化处理。以融合文献和作者的关键词共现矩阵 $ATR^{m\times m}$ 为例，记矩阵 $ATR^{m\times m}$ 第 i 行第 j 列的元素为 atr_{ij}，按公式（4）对其进行相似化处理后得到矩阵 $ATR'^{m\times m}$。

$$atr'_{ij}=\frac{atr_{ij}^2}{atr_{ii}*atr_{jj}}\ (i,j=1,2,\cdots,m) \qquad 公式（4）$$

（2）非对称视角下的标准化处理。上述方法是在对称视角下对关键词共现矩阵进行了标准化处理，虽然两个关键词的共现频次是唯一的，但是受单个关键词出现频次的影响，高频关键词与很多词存在关联，而低频词只与少数词存在关联，因此从高频词视角下计算的关联度与从低频词视角下计算的关联度是不同的，本文考虑利用公式（5）对矩阵 $ATR^{m\times m}$ 进行非对称视角下的相似性度量得到矩阵 $ATR''^{m\times m}$。

$$atr''_{ij}=\frac{atr_{ij}}{atr_{ii}}\quad(i,j=1,2,\cdots,m)\qquad \text{公式（5）}$$

4.2 研究方法

4.2.1 非负矩阵分解

非负矩阵分解起源于主成分分析，最早由 P. Paatero 等 [23] 提出，被称为正矩阵分解，其基本思想是将一个非负的矩阵分解为左右两个非负矩阵的乘积。对于关键词共现矩阵 $A\in R_{+}^{m\times m}$ 来说，m 表示关键词数，利用上述介绍的非负矩阵分解算法将其分解为 $A^{m\times m}\approx U^{m\times r}*V^{r\times m}$，其中矩阵 $V^{r\times m}$ 的行可以解释为 r 个主题，每行元素表示为词表中 m 个关键词在该主题中的非负权重，因此可以对词表的每一行按权重值大小进行排列，从而得到每个主题所包含的关键词种类，并根据关键词的权重值大小对主题进行命名 [24]。

4.2.2 非负张量分解

张量是一个多维数组，最常用的张量分解方法有 CP 分解和 Tucker 分解 [25]。CP 分解是将一个 n 阶张量分解成多个秩为 1 的张量的和的形式 [26]，Tucker 分解则是将其分解成一个核心张量与若干个因子矩阵乘积的形式，核心张量可以看成原张量的浓缩形式 [27]，当核心张量是一个对角的张量时，Tucker 分解则退化成了 CP 分解 [28-30]（见图 2）。非负张量分解则是非负矩阵分解在高维空间中的拓展，它既保留了张量的优点，又避免了负元素的出现，被广泛应用于图像处理、音频分类文本挖掘等领域。

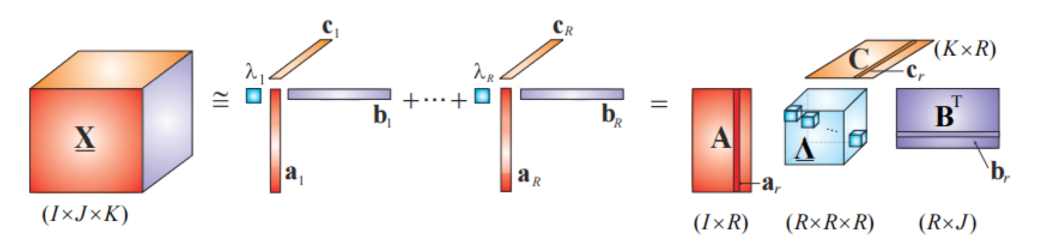


图 2 三阶张量的 CP 分解与 Tucker 分解

在利用非负张量分解进行主题识别时，首先需要构建一个合适的张量，以三阶张量为例，由于本文是基于关键词共现频次变化视角进行动态主题识别，因此本文构建了 < 关键词，关键词，年份 > 的三阶张量 $X^{I\times I\times K}$，如图 3 所示，其中关键词共现矩阵中的黑色圆圈代表关键词之间的共现强度，对该张量进行非负张量分解便可得到因子矩阵 $A^{I\times R}$、$B^{R\times I}$、$C^{K\times R}$，以及核心张量 $\Lambda^{R\times R\times R}$，其中 I 代表关键词种类数，K 代表年份数，R 代表聚类个数，与非负矩阵分解算法结果类似，非负张量分解算法中的因子矩阵 $A^{I\times R}$、$B^{R\times I}$ 均可解释为 R 个主题以及每个主题下包含的关键词种类及权重值大小，且两个因子矩阵下的聚类结果一致，此外因子矩阵 $C^{K\times R}$ 还可解释为 R 个主题在各个年份所占的权重值即主题研究热度，核心张量 $\Lambda^{R\times R\times R}$ 则可解释为 R 个主题的综合强度，由此便将 < 关键词，关键词，年份 > 的三阶张量降维成了 < 主题，年份 > 的二阶矩阵，从而可以进行主题的动态识别，如图 3 所示，主题框中的黑色圆圈大小代表主题在该年份所出现的强度大小。

知识管理论坛，2022（2）:197-208
DOI: 10.13266/j.issn.2095-5472.2022.017

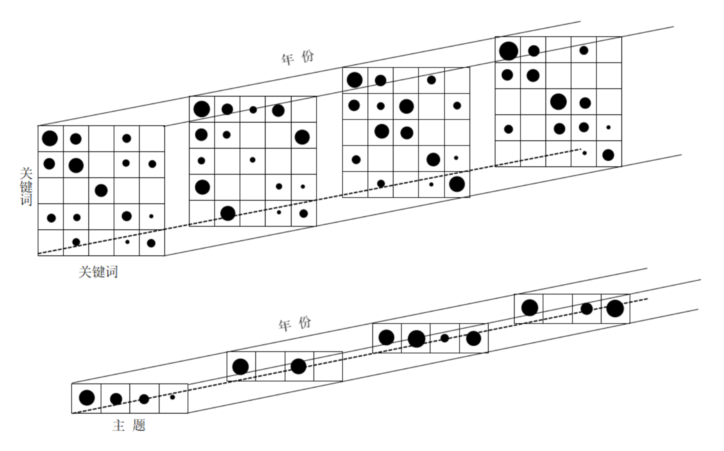


图 3 基于非负张量分解算法的动态主题识别示意图

5 结果分析

基于词共现频次变化视角进行动态主题识别时，首先需要构造合适的词共现矩阵，因此本文首先对几种词共现矩阵的构造方式及数据处理方法进行对比，然后选择合适的方法进行张量的构造与动态主题的识别。本文首先进行了两组对照实验，第一组实验对比了基于文献的关键词共现矩阵、基于作者的关键词共现矩阵以及融合文献与作者的关键词共现矩阵在主题识别结果上的差异；第二组实验在第一组实验结果的基础上，选择一种数据集构建方式，对比了共现矩阵进行相似化处理操作对主题识别结果的影响。

5.1 数据集构建组实验结果分析

通过多次实验发现：当类团数多于 5 类时，会出现部分类团中的关键词高度重叠的情形，因此本文将类团数定为 5 类，三种关键词共现矩阵下的非负矩阵分解聚类结果见表 1。可以看出，非负矩阵分解算法下的聚类结果中各类团里的关键词权重值大小均非负，弥补了主成分分析中权重值可正可负的不足，各类团中的关键词种类也有重复，弥补了系统聚类法中一个关键词只属于一个类团的不足，与现实情况相吻合。具体来看，三种关键词共现矩阵下的聚类结果既存在相同之处也呈现出差异：

首先，三种关键词共现矩阵下每个类团中的主导词（权重值最高的关键词）基本一致，这些主导词可以辅助于类团的命名，由此说明不管是在文献视角下还是作者视角下，国外知识管理领域近 5 年的研究热点基本相同，主要有 Knowledge Sharing、Innovation、Intellectual capital、Knowledge、Organizational performance、SEMs 等；不同之处在于每个大主题下的研究方向有所差异（即每个类团中权重值低的关键词种类有所差异），如文献视角下的 Innovation 主题中的关键词按权重值排序依次为 SMEs、Performance、Dynamic capabilities、Entrepreneurship 等，作者视角下 Innovation 主题中的关键词按权重值排序依次为 SMEs、Dynamic capabilities、Organizational performance、Information technology 等，两种视角下的创新主题研究都聚焦于企业，但文献视角下的企业创新侧重于企业家精神，而作者视角下的企业创新侧重于信息技术。

此外，通过 jaccard 相似度算法计算出每种聚类结果下各主题之间的关联度，得到关联度均值、极差和标准差等统计数据（图 4- 图 6）。可以看出，基于文献视角的聚类结果中每个主题与该聚类结果下其他主题的关联度均值都是最高，且极差和标准差最小；基于作者视角的聚类结果中每个主题与该聚类结果下其他主题的关联度均值都比较低，且极差和标准差都较大；而融合两种视角下关键词共现矩阵的聚类结果的主题关联度统计数据介于单视角结果之

表 1 三种关键词共现矩阵聚类结果

基于文献的关键词共现矩阵聚类结果

主题1		主题2		主题3		主题4		主题5	
关键词	权重	关键词	权重	关键词	权重	关键词	权重	关键词	权重
Knowledge Sharing	0.975	Innovation	0.973	Intellectual capital	0.930	Knowledge	0.908	Organizational performance	0.631
Knowledge creation	0.085	SMEs	0.093	Human capital	0.211	Knowledge transfer	0.235	Organizational learning	0.566
trust	0.055	Performance	0.067	Relational capital	0.130	Management	0.145	SMEs	0.202
Knowledge transfer	0.055	Absorptive capacity	0.048	structural capital	0.097	Higher education	0.098	structural equation modeling	0.188
Tacit knowledge	0.049	entrepreneurship	0.045	Performance	0.090	case study	0.089	Knowledge creation	0.183

基于作者的关键词共现矩阵聚类结果

主题1		主题2		主题3		主题4		主题5	
关键词	权重	关键词	权重	关键词	权重	关键词	权重	关键词	权重
Knowledge Sharing	0.961	Innovation	0.963	Intellectual capital	0.886	Organizational performance	0.594	SMEs	0.513
Knowledge creation	0.109	SMEs	0.121	Human capital	0.226	knowledge management processes	0.498	Open innovation	0.485
Knowledge transfer	0.088	Dynamic capabilities	0.078	Innovation performance	0.186	Knowledge	0.257	Innovation performance	0.314
Tacit knowledge	0.062	Organizational performance	0.051	Relational capital	0.153	Organizational learning	0.239	Absorptive capacity	0.244
trust	0.055	information technology	0.050	Organizational learning	0.108	structural equation modeling	0.199	Big data	0.243

基于文献和作者的关键词共现矩阵聚类结果

主题1		主题2		主题3		主题4		主题5	
关键词	权重	关键词	权重	关键词	权重	关键词	权重	关键词	权重
Knowledge Sharing	0.970	Innovation	0.971	Intellectual capital	0.908	Organizational performance	0.623	Knowledge	0.885
Knowledge creation	0.097	SMEs	0.111	Human capital	0.220	Organizational learning	0.411	Knowledge transfer	0.181
Knowledge transfer	0.074	Performance	0.059	Innovation performance	0.159	knowledge management processes	0.289	Management	0.168
trust	0.056	Dynamic capabilities	0.049	Relational capital	0.144	SMEs	0.229	Open innovation	0.118
Tacit knowledge	0.055	entrepreneurship	0.047	structural capital	0.101	Dynamic capabilities	0.212	Big data	0.115

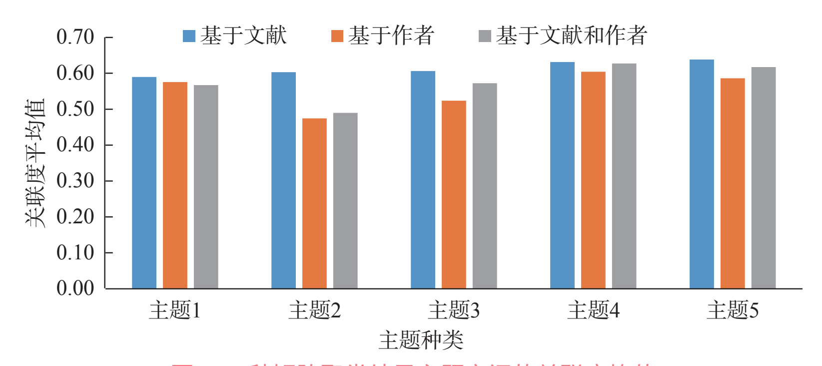


图 4 三种矩阵聚类结果主题之间的关联度均值

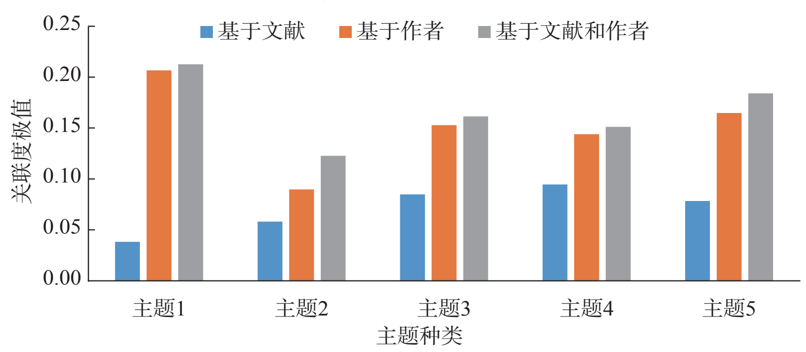


图 5 三种矩阵聚类结果主题之间的关联度极差

间。由此说明，作者视角下的聚类结果中各主题之间的区分度比文献视角下的聚类结果主题区分度更为明显，这是由于文献数量远多于作者数量，文献视角下的聚类结果可以对领域主题进行深入的挖掘，而作者视角下的聚类结果可以对领域主题进行全面的识别。结合三种聚类结果下各主题所包含的关键词个数（见图 7）可知，文献视角下的每个主题所包含的关键词种类较作者视角下的关键词种类更多，即主题内容挖掘得更为深入细致。因此，融合了文献和作者的关键词共现矩阵相较于单一视角下的关键词共现矩阵聚类结果既能全面地反映领域内的研究情况，又能对研究内容进行深入细致的挖掘。

ChinaXiv合作期刊
知识管理论坛，2022（2）:197-208
DOI: 10.13266/j.issn.2095-5472.2022.017

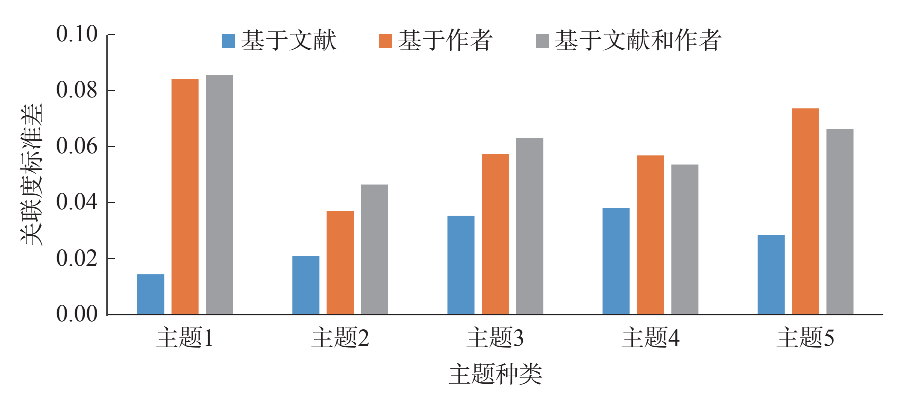


图 6 三种矩阵聚类结果主题之间的关联度标准差

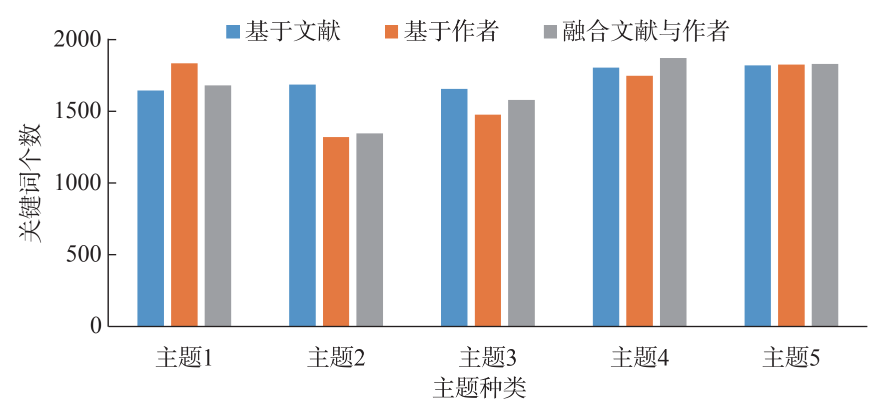


图 7 三种矩阵聚类结果各主题的关键词种类数

该组实验结果表明：文献是新知识、新技术的载体，代表了一个领域的最新研究成果，随着知识大爆炸时代的来临，文献数量迅猛增长，基于文献的关键词共现矩阵聚类结果可以表征一个领域内的热门研究主题与研究前沿，且由于文献数量远远多于作者数量，文献视角下的关键词共现矩阵可以对领域内的研究情况进行更为细致深入的挖掘；而作者则是长期耕耘在某一研究方向上的创造者，基于作者的关键词共现矩阵聚类结果可以表征领域内的经典研究主题，且对领域内的研究情况进行全面的反映。融合了文献和作者的关键词共现矩阵的聚类结果既能全面又能深入细致地反映领域内的研究情况。

5.2 数据集处理组实验结果分析

第一组实验结果表明：基于融合文献和作者双视角的关键词共现矩阵的主题识别结果能更好地反映领域内的研究情况，因此本文以该矩阵为例继续进行下一步分析。首先对融合文献和作者双视角下的关键词共现矩阵在对称视角下和非对称视角下进行标准化处理，然后利用非负矩阵分解算法对经标准化操作处理前后的关键词共现矩阵进行聚类，聚类结果见表 2。

可以看出，未经标准化处理的共现矩阵聚类结果与在非对称视角下进行标准化处理的共现矩阵聚类结果存在部分主题的主导词相同的情况（如 Knowledge sharing、Innovation、Knowledge 等），而在对称视角下进行标准化处理的共现矩阵聚类结果则差异较大，通过查看原始数据发现，未经标准化操作和在非对称视角下进行标准化操作的聚类结果中各主题下的主导词一般为高频关键词，且类团中的关键词权重值差异明显，而在对称视角下进行标准化操作的聚类结果中各主题下的关键词出现的频次都比较低，且各类团中的关键词权重差异不

表 2 相似化处理前后聚类结果

原始关键词共现矩阵聚类结果									
主题1		主题2		主题3		主题4		主题5	
关键词	权重	关键词	权重	关键词	权重	关键词	权重	关键词	权重
Knowledge Sharing	0.970	Innovation	0.971	Intellectual capital	0.908	Organizational performance	0.623	Knowledge	0.885
Knowledge creation	0.097	SMEs	0.111	Human capital	0.220	Organizational learning	0.411	Knowledge transfer	0.181
Knowledge transfer	0.074	Performance	0.059	Innovation performance	0.159	knowledge management processes	0.289	Management	0.168
trust	0.056	Dynamic capabilities	0.049	Relational capital	0.144	SMEs	0.229	Open innovation	0.118
Tacit knowledge	0.055	entrepreneurship	0.047	structural capital	0.101	Dynamic capabilities	0.212	Big data	0.115
对称相似化处理关键词共现矩阵聚类结果									
主题1		主题2		主题3		主题4		主题5	
关键词	权重	关键词	权重	关键词	权重	关键词	权重	关键词	权重
Information flow analysis	0.448	urban governance	0.470	transport technologies	0.499	Case management	0.386	Code switching	0.412
Alumni collaboration	0.448	governance configuration	0.470	capacity mapping	0.499	Process-oriented knowledge management	0.373	Corporate language	0.412
University alumni	0.448	relational approach	0.470	horizon scanning	0.499	Dynamic business process management	0.358	Language diversity	0.394
Data flow	0.448	uncertainty	0.326	transport innovation	0.290	ambulatory care information systems	0.311	multinational organisations	0.248
agricultural ontology	0.157	transition	0.326	transport research	0.290	testing and evaluation of health information technology	0.311	climate	0.229
非对称相似化处理关键词共现矩阵聚类结果									
主题1		主题2		主题3		主题4		主题5	
关键词	权重	关键词	权重	关键词	权重	关键词	权重	关键词	权重
Knowledge Sharing	0.945	Innovation	0.888	ontology	0.557	Knowledge	0.541	Sustainability	0.560
Language diversity	0.070	SMEs	0.198	Intellectual capital	0.464	Absorptive capacity	0.218	complexity	0.339
Knowledge creation	0.070	Organizational performance	0.091	Semantic web	0.299	Knowledge management	0.201	uncertainty	0.321
trust	0.067	information technology	0.078	Knowledge management systems	0.163	structural equation modeling	0.196	transition	0.317
Code switching	0.057	Performance	0.069	Linked Open Data	0.155	Higher education	0.188	urban governance	0.286

大，这是因为对称视角下的标准化可以消除高频关键词的影响。此外，在非对称视角下的标准化处理操作后的聚类结果除了将高频关键词聚拢，也将一部分低频关键词进行聚拢，这是由于一些关键词虽然出现的频次不高，但是每一次出现都伴随着其他词一起出现，这些词的关联度非常高，因而被聚为一类，而其他两种聚类结果则不具这一特点。

该组实验结果表明：使用原始关键词共现矩阵或对其进行非对称视角下的标准化处理，可以分析领域内的热点研究主题，因为高频关键词往往能代表某一领域的研究重点与热点，其中经非对称标准化处理后的关键词共现矩阵聚类结果除了可以研究高频关键词的类团，也涵盖了低频关键词的聚拢情况，可以更加全面地分析领域内的研究情况。使用对称视角下标准化处理的关键词共现矩阵可以分析领域内的最新前沿研究动向，在对称视角下进行标准化处理后的聚类结果既消除了高频关键词的影响，也未割除关键词之间的关联性。

5.3 动态主题识别结果分析

基于前两组的实验结果，第三组实验仍以融合了文献和作者双视角下的关键词共现矩阵数据为例，并进行非对称视角下的标准化处理操作，然后对比非负矩阵分解算法和非负张量分解算法在动态主题识别过程中的优劣性。由于非负矩阵分解算法处理的数据是矩阵形式，因此需要对 2017-2021 年期间的关键词共现矩阵按年进行时间分片，共需进行 5 次聚类，每年聚类的数据集为当年出现的所有关键词之间的共现矩阵；非负张量分解算法可以处理高维数据形式，因此可以直接对 2017-2021 年的所有关键词进行整体聚类，首先构造一个三阶张量，按年份维度可划分为 5 片，每片为 2017-2021 年期间出现的所有关键词在某一年份中的共现矩阵。非负矩阵分解算法和非负张量分解算法的聚类结果见表 3。

可以看出，非负矩阵分解算法下的聚类结果，在 2017-2021 年期间各年份的主要研究热点大致相同（每个类团中的主导关键词大致相同），但每个研究热点下的研究方向与研究细度略有差异（每个类团中的关键词数量及种类有所差异），而非负张量分解只对 2017-2021 年期间的关键词进行了一次聚类，聚类结果与非负矩阵分解算法的结果整体较为吻合（非负张量分解的聚类结果中的各主导词为非负矩阵分解聚类结果 5 年内出现较多的主导词）。

非负矩阵分解算法对 2017-2021 年期间的关键词共现矩阵进行了逐年多次聚类，而非负张量分解算法则是利用五年间关键词联系及演化得到五年间主题的识别与演化，即它所聚类出的主题为这 5 年间出现的所有主题，然后利

知识管理论坛，2022（2）:197-208
DOI: 10.13266/j.issn.2095-5472.2022.017

表 3 2017-2021 年期间两种聚类算法结果对比

主题1 关键词	权重	主题2 关键词	权重	主题3 关键词	权重	主题4 关键词	权重	主题5 关键词	权重
2017年关键词聚类结果（非负矩阵分解）									
Knowledge Sharing	0.931	Innovation	0.933	Intellectual capital	0.877	Knowledge	0.869	ontology	0.785
Knowledge transfer	0.124	SMEs	0.216	Human capital	0.187	Management	0.280	Knowledge management systems	0.424
Knowledge management systems	0.115	case study	0.075	Relational capital	0.116	Big data	0.104	Semantic web	0.241
2018年关键词聚类结果（非负矩阵分解）									
Knowledge Sharing	0.953	Innovation	0.948	Intellectual capital	0.868	Knowledge	0.889	Absorptive capacity	0.521
Organizational culture	0.089	information technology	0.117	Human capital	0.231	Management	0.151	ontology	0.434
Organizational learning	0.076	entrepreneurship	0.075	Innovation performance	0.182	case study	0.117	Organizational performance	0.368
2019年关键词聚类结果（非负矩阵分解）									
Knowledge Sharing	0.968	Innovation	0.955	Knowledge management	0.471	SMEs	0.698	Absorptive capacity	0.594
Knowledge creation	0.109	Sustainability	0.080	Dynamic business process management	0.453	Knowledge creation	0.318	Intellectual capital	0.442
trust	0.048	Innovation performance	0.074	Case management	0.352	Organizational performance	0.223	Dynamic capabilities	0.418
2020年关键词聚类结果（非负矩阵分解）									
Innovation	0.936	Knowledge Sharing	0.935	Organizational performance	0.512	SMEs	0.898	Knowledge transfer	0.910
Organizational performance	0.099	Higher education	0.114	structural equation modeling	0.468	Thailand	0.129	Organizational learning	0.206
Performance	0.098	Knowledge creation	0.110	Organizational learning	0.409	Knowledge	0.125	Knowledge creation	0.136
2021年关键词聚类结果（非负矩阵分解）									
Innovation	0.925	Knowledge Sharing	0.902	Performance	0.663	Industry 4	0.457	Dynamic capabilities	0.607
Knowledge	0.140	Collaboration	0.142	Intellectual capital	0.317	Tacit knowledge	0.331	Innovation capability	0.519
Dynamic capabilities	0.083	Social Media	0.140	Organization	0.259	information technology	0.295	Conditions of knowledge management	0.203
2017-2021年关键词聚类结果（非负张量分解）									
Knowledge Sharing	0.976	Innovation	0.971	Intellectual capital	0.797	Knowledge	0.927	Knowledge creation	0.678
Knowledge transfer	0.068	SMEs	0.112	Organizational learning	0.316	Management	0.140	Absorptive capacity	0.478
trust	0.064	Organizational performance	0.070	Organizational performance	0.263	Big data	0.109	Knowledge Sharing	0.292

用分解后核心张量的结果，得到这所有主题在每年出现的概率或是研究强度，从而实现了只需一次聚类便可进行分析多年研究情况的动态主题识别。但是由于非负张量分解只进行了一次聚类，所以各年份出现的相同主题的研究内容都保持不变，相对综合，而非负矩阵分解是对各年分别进行单独聚类，因此不同年份可能主题相似，但内容有所差异，即非负矩阵分解在动态主题识别时对各主题的研究内容刻画得更为细致。

此外，通过对非负矩阵分解下的各年份聚类结果利用jaccard相似度算法计算主题相似度，得到主题演化脉络图（见图8），而非负张量分解下的聚类结果可以利用核心张量得到各年份主题的研究强度图（见图9），这种主题研究强度并非以主题的关键词数量或者频次来衡量，而是通过各年份关键词之间的共现变化关系而得出的主题演化强度，非负矩阵分解则较难实现这点。

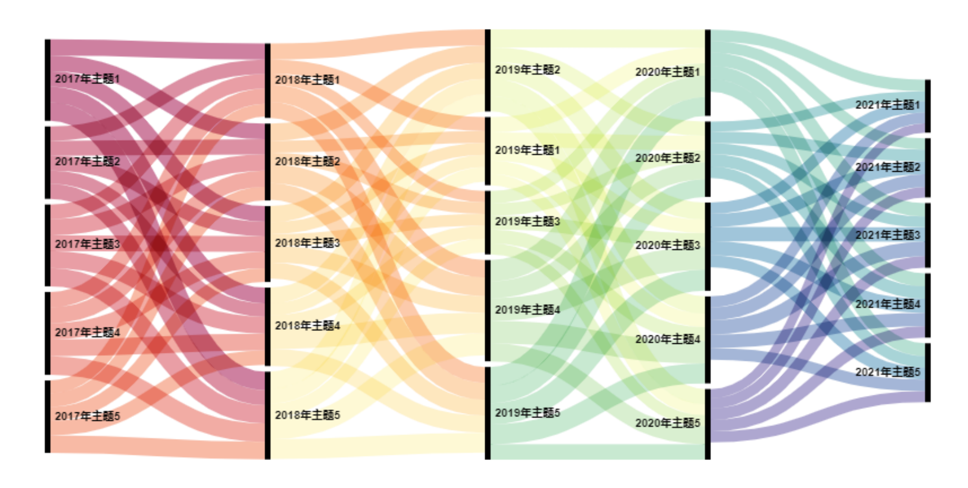


图 8 2017-2021 年知识管理领域主题演化（非负矩阵分解）

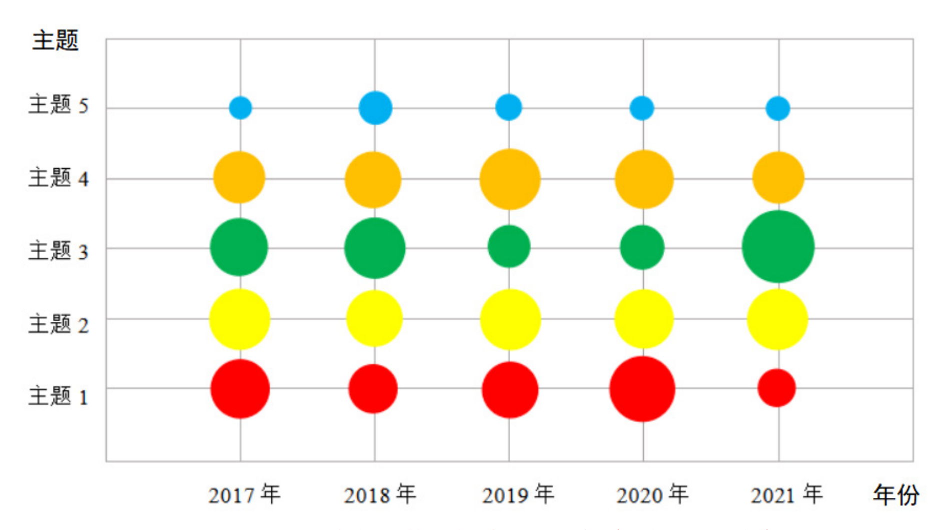


图 9 2017-2021 年知识管理领域主题强度（非负张量分解）

该组实验结果表明：如果想对领域内的研究情况进行大致的分析，可以采用非负张量分解算法，该算法简单快捷，只需一次聚类便可得到各年份的研究主题及研究强度等信息，大大降低了算法的复杂度，也减少了信息的损失。如果想细致地分析领域内各年份的研究情况可以采用非负矩阵分解进行逐年分析，这样可以得到各年份主题的具体研究内容及变化，也可以得到不同年份之间的主题演化情况，不足之处在于需要进行多次聚类及数据处理，且难以观察由于关键词共现演化带来的主题演化情况。

综上，在利用关键词共现数据进行领域主题识别时，选择融合文献和作者双视角下的关键词共现矩阵数据更能全面地反映领域内的研究情况；在利用共现数据进行主题识别时，需要对共现数据利用相似度度量进行标准化处理，其中在对称视角下进行标准化处理可以消除高频关键词的影响，分析领域内的前沿动向，在非对称视角下进行标准化处理可以研究领域内的热点问题；在进行动态主题识别过程中，非负张量分解算法可以简单快速地获取领域内的研究主题及其在各年的研究强度，而非负矩阵分解则可以更为细致深入地刻画主题以及主题的演化脉络，但是需要进行多次操作。

6 结束语

本文针对传统基于词共现矩阵的动态主题识别研究中需要进行多次聚类的不足，提出一种新的数据构建方式及处理方法，基于张量结构的数据形式可以在词共现矩阵中融入时间维度，尽可能地保留数据的原始信息，基于非负张量分解算法的动态主题识别只需进行一次聚类便可得到各年份的主题情况，有效避免了信息的损失。此外，本文还对几种词共现矩阵的构造方式及矩阵处理方法进行了探讨：在数据集的构建方式上，分别从文献视角、作者视角以及融合文献和作者双视角构建了关键词共现矩阵；在数据处理方式上，分别从对称视角和非对称视角利用相似性度量对共现矩阵进行了标准化操作，并对比了标准化操作对主题识别结果的影响。实验结果表明：融合文献和作者双视角下的关键词共现矩阵可以更全面地反映领域内的知识结构，对称视角下的标准化处理与非对称视角下的标准化处理在分析研究热点与研究前沿上各具优势。本文旨在为基于关键词共现的主题识别研究提供一些方法和流程上的参考，提高主题识别精度，为科技决策提供更好的支撑。

ChinaXiv合作期刊
知识管理论坛，2022（2）:197-208
DOI: 10.13266/j.issn.2095-5472.2022.017

参考文献：

[1] BUSH V. As we may think[J]. The Atlantic monthly, 1945 (7): 1-2 .

[2] 刘向 , 马费成 , 陈潇俊 , 等 . 知识网络的结构与演化——概念与理论进展 [J]. 情报科学 , 2011, 29(6): 801-809.

[3] 巴志超 , 杨子江 , 朱世伟 , 等 . 基于关键词语义网络的领域主题演化分析方法研究 [J]. 情报理论与实践 , 2016, 39(3): 67-72.

[4] 王莉亚 . 主题演化研究进展 [J]. 情报探索 , 2014(4): 29-32.

[5] 邵作运 , 李秀霞 . 引文分析法与内容分析法结合的文献知识发现方法综述 [J]. 情报理论与实践 , 2020, 43(3): 153-159.

[6] 邹丽雪 , 王丽 , 刘细文 . 利用引文构建的主题模型研究进展 [J]. 图书情报工作 , 2019, 63(23): 131-138.

[7] 祝清松 , 冷伏海 . 基于引文主路径文献共被引的主题演化分析 [J]. 情报学报 , 2014, 33(5): 498-506.

[8] 黄福 , 侯海燕 , 任佩丽 , 等 . 基于共被引与文献耦合的研究前沿探测方法鄰选 [J]. 情报杂志 , 2018, 37(12): 13-19, 35.

[9] 宋艳辉 , 武夷山 . 基于作者文献耦合分析的情报学知识结构研究 [J]. 图书情报工作 , 2014, 58(1): 117-123.

[10] 张洁 , 王红 . 基于词频分析和可视化共词网络图的国内外移动学习研究热点对比分析 [J]. 现代远距离教育 , 2014(2): 76-83.

[11] 叶春蕾 , 冷伏海 . 基于共词分析的学科主题演化方法改进研究 [J]. 情报理论与实践 , 2012, 35(3): 79-82.

[12] 奉国和 , 孔泳欣 . 基于时间加权关键词词频分析的学科热点研究 [J]. 情报学报 , 2020, 39(1): 100-110.

[13] 储节旺 , 钱倩 . 基于词频分析的近 10 年知识管理的研究热点及研究方法 [J]. 情报科学 , 2014, 32(10): 156-160.

[14] 姜鑫 , 王德庄 , 马海群 . 关键词词频变化视角下我国"科学数据"领域研究主题演化分析 [J]. 现代情报 , 2018, 38(1): 141-146, 161.

[15] 赵丽梅 , 张花 . 我国大数据时代数字图书馆研究前沿分析——基于共词分析的视角 [J]. 情报科学 , 2019, 37(3): 97-104.

[16] 唐果媛 , 张薇 . 基于共词分析法的学科主题演化研究进展与分析 [J]. 图书情报工作 , 2015, 59(5): 128-136.

[17] 胡吉明 , 陈果 . 基于动态 LDA 主题模型的内容主题挖掘与演化 [J]. 图书情报工作 , 2014, 58(2): 138-142.

[18] 杨超 , 朱东华 , 汪雪锋 , 等 . 专利技术主题分析 : 基于 SAO 结构的 LDA 主题模型方法 [J]. 图书情报工作 , 2017, 61(3): 86-96.

[19] KIM J, HWANG M, JEONG D H, et al. Technology trends analysis and forecasting application based on decision tree and statistical feature analysis[J]. Expert systems with applications, 2012, 39(16): 12618-12625.

[20] WALTMAN L, VANECK N J. Some comments on the question whether co-occurrence data should be normalized[J]. Journal of the American Society for Information Science and Technology, 2007, 58(11): 1701-1703.

[21] LEYDESDORFF L. Should co-occurrence data be normalized? a rejoinder[J]. Journal of the American Society for Information Science and Technology, 2007, 58(14): 2411-2413.

[22] van ECK N J, WALTMAN L. How to normalize cooccurrence data? an analysis of some well-known similarity measures[J].Journal of the American Society for Information Science and Technology, 2009, 60(8): 1635-1651.

[23] PAATERO P, TAPPER U. Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values[J]. Environmetrics, 1994, 5(2): 111-126.

[24] 章祥荪 , 张忠元 . 非负矩阵分解 : 模型、算法和应用 [J]. 重庆师范大学学报 (自然科学版), 2013, 30(6): 1-8.

[25] 吴继冰 , 黄宏斌 , 邓苏 . 网络异构信息的张量分解聚类方法 [J]. 国防科技大学学报 , 2018, 40(5): 146-152, 170.

[26] 熊李艳 , 何雄 , 黄晓辉 , 等 . 张量分解算法研究与应用综述 [J]. 华东交通大学学报 , 2018, 35(2): 120-128.

[27] 程齐凯 , 王晓光 . 一种基于共词网络社区的科研主题演化分析框架 [J]. 图书情报工作 , 2013, 57(8): 91-96.

[28] LUO J, GWUN O. A comparison of sift PCA-SIFT and SURF[J]. International journal of image processing, 2009, 3(4): 143-152.

[29] CICHOCKI A, ZDUNEK R, PHAN A H, et al. Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation[M]. Hoboken: Wiley Publishing, 2009.

[30] 熊李艳 , 何雄 , 黄晓辉 , 等 . 张量分解算法研究与应用综述 [J]. 华东交通大学学报 , 2018, 35(2): 120-128.

作者贡献说明：

方　洁：提出研究思路，进行论文指导及修订；

崔兰兰：进行数据采集、研究思路设计、数据分析、论文撰写及修订。

Research on Dynamic Topic Recognition Based on the Change of Word Co-Occurrence Frequency

Xi Chongjun Liu Wenbin Ding Kai
Institute of Science and Technology Information of China, Beijing 100038

Abstract: [Purpose/Significance] The research on topic recognition is very important to clarify the knowledge structure and research hotspots in the field. Dynamic identification of domain topics can help researchers understand and master the development trend and future trend of the field. **[Method/Process]** Using the data structure form of tensor, this paper integrated the time dimension into the word co-occurrence matrix, and only needed one clustering to identify the dynamic topic. **[Result/Conclusion]** Tensor structure and non-negative tensor decomposition algorithm provide a new method for dynamic topic recognition from the perspective of word co-occurrence frequency change. Compared with traditional methods, this method is simpler and faster, and effectively avoids the loss of information.

Keywords: keyword co-occurrence　non-negative matrix factorization　non-negative tensor factorization　dynamic topic recognition　knowledge management